WORKLOAD MEASUREMENT IN SYSTEM DESIGN AND EVALUATION

F. Thomas Eggemeier
WRIGHT STATE UNIVERSITY
Dayton, Ohio

Clark A. Shingledecker and Mark S. Crabtree
ERGOMETRICS TECHNOLOGY, INC.
Dayton, Ohio

ABSTRACT

Because of its central role in system development, workload measurement has been extensively researched. These efforts have produced a variety of workload assessment techniques, many of which can be classified as either subjective, physiological, or behavioral measures. These categories of measures can vary along several dimensions that can be used as criteria in selection of a technique for a particular application. The proposed selection criteria include the sensitivity, diagnosticity, and intrusiveness associated with a technique. Different stages of system design can require techniques that differ on the noted dimensions. Since no technique is capable of meeting all of the applicable criteria, a comprehensive approach to workload assessment will require a battery of subjective, physiological, and behavioral measures. Future research dealing with comparative evaluation of the various assessment techniques along the noted dimensions will be required in order to refine workload metric selection criteria.

INTRODUCTION

The system development process consists of a series of stages which range from conceptual development through test and evaluation of the system. Although a variety of human factors engineering functions (e.g., control/display design; function allocation) are performed during the various stages, the primary purpose of many of these functions is to ensure that system demands do not exceed the information processing capabilities of the operator. Processing overload represents a major factor that can contribute to decrements in operator performance and to degradations in system effectiveness.

Workload assessment techniques are principally designed to measure the degree of operator processing capacity which is expended in performing a particular task or system function. By measuring expended capacity, existing or potential processing overloads may be identified and breakdowns in operator performance avoided. Adequate workload assessment procedures are therefore critical to many of the human factors functions that are performed throughout the design process.

Because of its central role during system design, workload measurement has been extensively researched in recent years. A wide variety of procedures have been proposed to measure workload, but most empirical techniques can be classified as belonging to one of three major categories: (1) subjective opinion measures (e.g., rating scales), (2) physiological techniques (e.g., evoked cortical potentials), and (3) behavioral measures. The behavioral category is typically divided into two major procedures, primary and secondary task techniques. Primary task procedures measure the adequacy of operator performance on the task or system of interest, while secondary task methodology indexes primary task capacity expenditure by assessing the operator's capability to perform a second concurrent task. All three major classes of measures have been employed with varying degrees of success in a variety of applications (O'Donnell & Eggemeier, 1985; Wierwille & Williges, 1978).

The availability of several classes of workload assessment techniques raises several issues for the system designer, one of which is the choice of the type of techniques(s) to be employed in a given application. The potential importance of this choice is underscored by the fact that when several measures of load have been applied together, they sometimes exhibit a pattern of dissociation (Eggemeier, Crabtree, & LaPointe, 1983; Wickens & Derrick, 1981; Wickens & Yeh, 1983; Wierwille & Casali, 1983). Current workload theory maintains that the noted dissociations may be at least partially attributable to two factors: (1) the nature of the relationship between levels of workload and the adequacy of operator performance, and (2) the characteristics of the processing capacity limitations within the human system. These approaches to dissociation suggest two criteria that can be considered in choice of an assessment technique. Both criteria are related to the type of question that is to be answered by the workload measure, and indicate that some techniques may be more appropriately applied than others in addressing particular objectives during system design. In addition to the specific objective to be addressed by the workload measure, an important factor related to the choice of an assessment technique deals with the potential of the procedure to disrupt on-

going primary task performance. Such disruptions can be more critical during some phases of design than others, and should be considered in choosing a technique. Taken together, these criteria provide some basis for evaluating the usefulness of a technique for a particular purpose. The following section outlines the suggested criteria in more detail, and illustrates their application to the major categories of assessment techniques.

WORKLOAD TECHNIQUE EVALUATION CRITERIA

A number of criteria applicable to workload assessment techniques have been proposed in the recent literature (e.g., Eggemeier, 1984; Shingledecker, 1983; Wierwille & Williges, 1978). Several of the criteria are related to the factors noted above, and include the (1) sensitivity, (2) diagnosticity, and (3) intrusiveness associated with various techniques.

Sensitivity

The sensitivity of a technique is determined by its capability to discriminate differences in loading imposed by a task or system function. At a very general level, sensitivity is related to the hypothetical function which relates workload levels to the adequacy of operator performance. This function can be characterized as consisting of at least two regions, one incorporating low to moderate levels of load, while the second spans higher levels of loading. In the first region, increases in workload are typically not accompanied by variations in performance, since the operator has sufficient spare processing capacity to compensate for such increases and maintain primary task performance. The second or higher workload region, on the other hand, is characterized by a monotonic relationship between load and performance, since it is assumed that spare processing capacity has been exhausted and the operator can no longer compensate for increased demand. The proposed hypothetical relationship suggests that primary task measures will be relatively insensitive to variations in loading levels in the first region, but will discriminate increases in workload in the second region where processing overloads exist. Other techniques (subjective measures, secondary task methodology, physiological measures) which are intended to index the degree of expended effort or processing capacity, should, however, be sensitive to loading levels in the first region where no overload exists.

An example of such differences in sensitivity between primary task measures and a subjective metric are illustrated in Figure 1, which is adapted from Eggemeier et al. (1983). Subjects in this experiment performed a short-term memory task at three interstimulus intervals, and rated the workload associated with each demand level using the Subjective Workload Assessment Technique (SWAT). As is clear from Figure 1, SWAT ratings varied substantially as a function of demand level, while no significant differences were obtained in the primary task measure of memory errors. This pattern of dissociation is consistent with the expectations of the hypothetical workload-performance relationship, and can be interpreted as suggesting that primary task performance was maintained at a cost of increased effort/capacity expenditure, which was reflected in the subjective measure. A similar example of dissociation was recently reported by Schifflet, Linton, and Spicuzza (1982), who employed a secondary memory search task to discriminate loading differences in two display options. These differences were not reflected in measures of the primary task performance associated with the options.

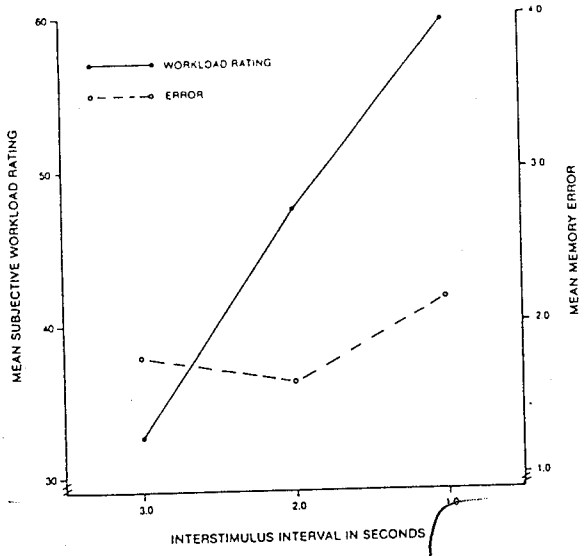

Figure 1. Mean Subjective Workload Rating and Memory Error as a Function of Interstimulus Interval (From: Eggemeier et al., 1983).

The pattern of dissociation predicted by the proposed workload-performance function suggests that the use of primary task procedures versus other metrics should be based on the objective that is to be addressed by the workload measure. If the objective is to determine if processing overloads that are associated with degraded performance actually exist, primary task measures should be employed. On the other hand, if the objective is to evaluate the potential for the overload between two design options (e.g., displays) that yield adequate primary task performance, then a potentially more sensitive metric (e.g., subjective, secondary task) should be employed. This objective would be important when the designer anticipates that other factors (e.g., environmental stressors) might contribute additional demand that

would be sufficient to overload the operator and cause degraded performance.

Diagnosticity

Diagnosticity refers to the capability of an assessment technique to discriminate differences in the loading imposed on specific processing capacities/resources within the human system. This criterion is based upon the multiple resources theory (e.g., Wickens, 1984) regarding the nature of operator processing capacity limitations. In essence, this theory proposes that the processing capacity expended during task performance is not unitary, but is drawn from several independent sources or pools, each with its own limited capacity/resources. An important implication of this approach is that it is possible to overload or fully expend the resources associated with one source, while not exhausting the resources of another pool. One current version (Wickens, 1984) of multiple resources theory maintains that perceptual and central processing functions draw on one resource pool, while motor output functions draw on another pool. Under this approach, a workload metric that was maximally sensitive to motor output capacity expenditure might not reflect variations in perceptual and central processing loading, and would be highly diagnostic of motor demand.

This type of diagnosticity has been demonstrated by a secondary interval production task in a series of experiments conducted by Shingledecker, Acton, and Crabtree (1983). The interval production task (IPT) was performed concurrently with three other tasks, including subcritical tracking (motor resources), memory search (central processing resources), and display monitoring (perceptual resources). Demand levels in the three tasks were manipulated by varying the level of instability (lambda) in the tracking task, the number of items to be searched in the memory task, and the number of displays and discriminability of signals in the monitoring task. As illustrated in Figure 2, IPT performance demonstrated a marked sensitivity to demand manipulations in the tracking task, but showed little or no sensitivity to similar manipulations in the memory task. The monitoring task results were essentially the same as those illustrated for the memory task, with the IPT failing to reflect any significant differences as a function of loading levels.

Similar patterns of diagnosticity has been shown with the event-related brain potential (Isreal, Chesney, Wickens, & Donchin, 1980), and with other secondary tasks (Wickens & Kessel, 1980). On the other hand, other physiological metrics such as pupil diameter (Beatty, 1982) and some workload rating scales (Reid, 1985) appear sensitive to a variety of different demands. These data suggest that subjective rating scales and some physiological measures are not particularly diagnostic, and

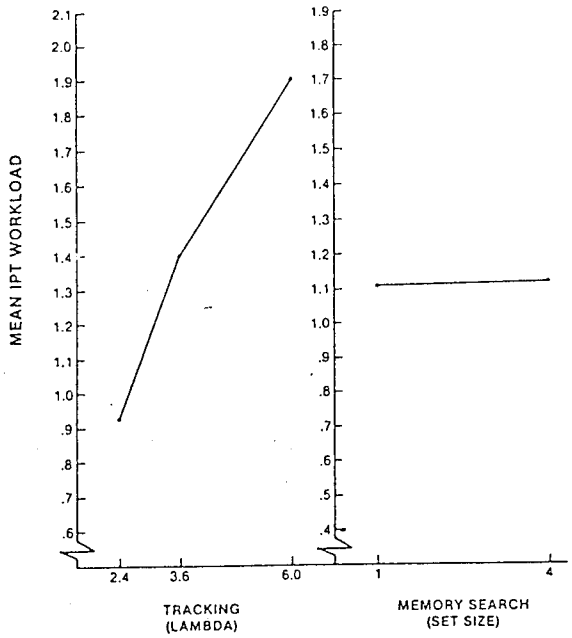

Figure 2. Mean Interval Production Task Performance as a Function of Difficulty Levels in Two Tasks (From: Shingledecker et al., 1983).

are sensitive to capacity expenditure throughout the system. Other physiological measures and some secondary tasks such as those noted above, however, appear to be highly diagnostic. Comparisons of workload indices from a generally sensitive metric with those from a highly diagnostic procedure could result in a pattern of dissociation, since the two techniques would be providing somewhat different information about the load imposed by particular system demands.

Such differences in diagnosticity suggest that different types of measures can play complementary roles during system design. Less diagnsotic measures could serve as initial screening devices to detect high levels of loading during any phase of task performance, while more diagnostic procedures could be subsequently used to identify the particular source (e.g., central processing vs. motor output) of any such overloads. Choice of an assessment technique on the basis of the diagnosticity criterion, would, therefore, be dependent on the objective to be met by measuring the workload.

Intrusiveness

Intrusiveness is a third important characteristic of workload assessment techniques, and refers to the tendency of a procedure to cause degradations in on-going primary task performance. Such intrusiveness can be undesirable, since a technique which disrupts primary task performance may not accurately reflect the

levels of load that would ordinarily by imposed by unimpaired performance. Significant levels of intrusion can therefore cause problems in interpreting the results generated by use of an assessment procedure.

The interpretation problems associated with intrusion may, however, be less serious in some instances than in others. For example, an investigator conducting a comparative evaluation of two display options might not be particularly concerned if the workload technique employed caused limited and equivalent decrements in the primary task performance associated with each option. Since the intent in this case is evaluation of the relative levels of load imposed by each option, valid conclusions could be drawn even though some intrusion was present. In these types of situations, a second set of considerations related to practical constraints imposed by the system design process may place additional limits on the degree of intrusiveness that is acceptable. The importance of degradations in primary task performance associated with intrusion can, for example, vary as a function of the stage of the design process which is under consideration. Levels of primary task intrusion that are acceptable in mockups or simulations that are typically associated with the earlier stages of design might not be acceptable during later in-flight test and evaluation of a prototype aircraft, due to the potential compromises in system safety involved. Choice of a workload measurement procedure on the basis of intrusiveness should therefore also take into account the practical aspects of the environment in which the measure is to be taken.

Despite the theoretical and practical importance of intrusion, the comparative data base related to the degree of intrusiveness that is typically associated with the various classes of workload assessment techniques is minimal. Although some steps toward establishment of such a data base have been taken recently (Shingledecker et al., 1983; Wierville & Casali, 1983), the data are not yet complete. Data from individual applications of each class of technique (O'Donnell & Eggemeier, 1985), however, suggest that secondary task measures may hold the greatest potential for intrusion. Subjective techniques that are typically applied at the completion of a task appear at present to minimize intrusion difficulties. Physiological measures that require no additional behavioral response on the part of the operator apparently also tend to limit the possibility of intrusion, although the use of some of the equipment (e.g., recording electrodes) required for these measures could pose a potential intrusion problem in some environments.

CONCLUSIONS

As is apparent from the foregoing discussion, none of the major classes of workload assessment techniques is totally capable of satisfying all of the objectives and meeting all of the constraints suggested by the three dimensions that have been proposed. Particular techniques are capable of meeting some criteria but not others, and a technique that is ideal for one application may not be as acceptable in another. For example, the requirement for a highly diagnostic assessment procedure for application in a simulation environment in which the potential for some intrusion does not represent a major practical problem could be met by a secondary task technique, while the need for a more globally sensitive measure in an operations environment that would not permit intrusion or extensive instrumentation might be more appropriately met by a subjective measurement procedure. Consequently, a comprehensive approach to workload assessment during system design currently requires the use of a battery of techniques, including subjective, physiological, and behavioral procedures.

Although some general guidelines for applications of workload assessment techniques during system design can be derived from current theory and data, additional comparative research is required in order to further refine the current bases for choosing a technique. It has already been noted that the comparative data base in intrusivenss is limited, and the same is true of the sensitivity and diagnosticity dimensions. For example, although it is assumed that application of subjective, physiological, or secondary task techniques can provide more sensitive indices of capacity expenditure under moderate demand levels than primary task measures, the data directly comparing these potentially more sensitive techniques under standard loading levels is very limited. Likewise, although there are instances in which the diagnosticity of particular secondary tasks has been demonstrated, these are also extremely limited and additional research is required to investigate the pattern of diagnosticity that might be associated with other tasks. Research efforts of these types that are related to the proposed dimensions should substantially contribute to the data base that can be used for choosing a workload assessment procedure for particular applications during system development.

ACKNOWLEDGEMENT

Preparation of this paper was partially supported by a contract with Wright State University from the Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio in conjunction with the Air Force Office of Scientific Research (Contract No. F33615-82-K-0522).

REFERENCES

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, 91, 276-292.

Eggemeier, F.T. (1984). Workload metrics for system evaluation. Proceedings of the Defense Research Group Panel VIII Workshop "Application of System Ergonomics to Weapon System Development", Shrivenham, England, pp. c/5-c/20.

Eggemeier, F.T., Crabtree, M.S., & LaPoint, P.A. (1983). The effect of delayed report on subjective ratings of mental workload. Proceedings of the Human Factors Society Twenty-Seventh Annual Meeting, pp. 139-143.

Isreal, J.B., Chesney, G.L., Wickens, C.D., and Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology, 17, 259-273.

O'Donnell, R.D., & Eggemeier, F.T. (1985). Workload assessment methodology. In L. Kaufman, J. Thomas, & K. Boff (Eds.), Handbook of Perception and Performance, New York: Wiley, in press.

Reid, G.B. (1985). Systematic development of a subjective measure of workload. Proceedings of the Ninth Congress of the International Ergonomics Association, in press.

Schifflet, S.G., Linton, P.M., & Spicuzza, R.J. (1982). Evaluation of a pilot workload assessment device to test alternative display formats and control handling qualities. Proceedings of the AIAA Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics, pp. 222-233.

Shingledecker, C.A. (1983). Behavioral and subjective workload metrics for operational environments. Proceedings of the AGARD Symposium on Sustained Intensive Air Operations, Physiological and Performance Aspects, (AGARD-CP-338), pp. 6/1-6/10.

Shingledecker, C.A., Acton, W.H., & Crabtree, M.S. Development and application of a criterion task set for workload metric evaluation. (Paper No. 831419). Warrendale, Pennsylvania: Society of Automotive Engineers Technical Paper.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & D.R. Davies (Eds.), Varieties of Attention, New York: Academic Press, 1984.

Wickens, C.D., & Derrick, W. (1981). Workload measurement and multiple resources. Proceedings of the IEEE Conference on Cybernetics and Society, pp. 600-603.

Wickens, C.D., & Kessel, C. (1980). The processing resource demands of failure detection in dynamic systems. Journal of Experimental Psychology: Human Perception and Performance, 6, 564-577.

Wickens, C.D., & Yeh, Y.Y. (1983). The dissociation of subjective ratings and performance: A multiple resources approach. Proceedings of the Human Factors Society Twenty-Seventh Annual Meeting, pp. 244-248.

Wierwille, W.W., & Casali, J.G. (1983). The sensitivity and intrusion of mental workload estimation techniques in piloting tasks. Blacksburg, Virginia: IEOR Department Report No: 8309, Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University.

Wierwille, W.W., & Williges, R.C. (1978). Survey and analysis of operator workload assessment techniques. Blacksburg, Virginia: Report No. S-78-101, Systemetrics, Inc.

ST#A AUTH: AL/CFHP (MR. REID-DSN 785-8749
PER TELECON, 6 JAN 95 CB